

RESEARCH

Open Access



SILVA, RDP, Greengenes, NCBI and OTT — how do these taxonomies compare?

Monika Balvočiūtė* and Daniel H. Huson

From The Fifteenth Asia Pacific Bioinformatics Conference
Shenzhen, China. 16–18 January 2017**Abstract****Background:** A key step in microbiome sequencing analysis is read assignment to taxonomic units. This is often performed using one of four taxonomic classifications, namely SILVA, RDP, Greengenes or NCBI. It is unclear how similar these are and how to compare analysis results that are based on different taxonomies.**Results:** We provide a method and software for mapping taxonomic entities from one taxonomy onto another. We use it to compare the four taxonomies and the Open Tree of life Taxonomy (OTT).**Conclusions:** While we find that SILVA, RDP and Greengenes map well into NCBI, and all four map well into the OTT, mapping the two larger taxonomies on to the smaller ones is problematic.**Keywords:** Metagenomics, Taxonomic classification, OTU assignment, NCBI, Silva, RDP, Greengenes, Open tree of life**Background**

Microbiome sequencing analysis is concerned with sequencing DNA from microorganisms living in certain environments without cultivating them in laboratory. In a typical taxonomy guided approach [1], sequencing reads are first binned into taxonomic units and then the microbial composition of samples is analyzed and compared in detail (see Fig. 1).

The two main technical ingredients of taxonomic analysis are the reference taxonomy used and the binning approach employed. Binning is usually performed either by aligning reads against reference sequences (e.g. [2]) or using k-mer based techniques (e.g. [3]). Taxonomic binning of 16S reads is usually based on one of these four taxonomies: SILVA [4], RDP [5], Greengenes [6] or NCBI [7].

How important is the choice of reference taxonomy, given the known inconsistencies of microbial classifications [8]? To address this, the aim of this paper is to determine how similar these four taxonomies are, and

whether results obtained using one classification can easily be carried over to another.

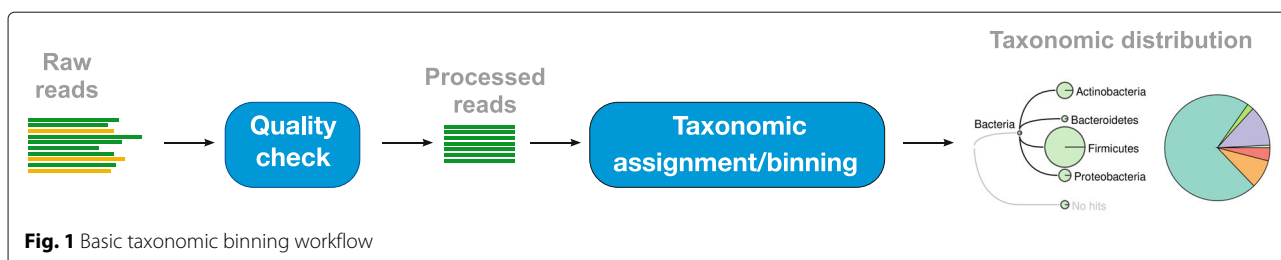
We define and explore an algorithm for mapping one taxonomy into another. This method allows us to compare taxonomies and is the basis for a tool that makes analyses on different classifications comparable to each other by mapping them onto a common taxonomy. While our main focus is on the four most popular taxonomic trees, we also consider the recently published Open Tree of life Taxonomy (OTT) [9].

We found that SILVA, RDP and Greengenes can be mapped into NCBI and OTT with few conflicts, but not vice versa. There is a great deal of difference between taxonomies that arise because of the differences in size and structure.

Taxonomic classifications

Each of the five taxonomies that we compare is based on a mixture of sources that have been compiled into taxonomies in different ways. They differ in both size and resolution (see Table 1). All taxonomies assign ranks to their nodes, the seven main ones being domain, phylum, class, order, family, genus and species. However, RDP and SILVA only go down to the genus level, whereas NCBI and

*Correspondence: monika.balvociete@uni-tuebingen.de
University of Tübingen, Department of Computer Science, Sand 14, 72076 Tübingen, Germany



OTT go down to the species level and below. The two latter taxonomies also have a number of intermediate ranks and contain many intermediate nodes (Fig. 2a). To simplify the comparison of taxonomies, we will consider only nodes associated with the seven main ranks.

Figure 2a shows the percentage of nodes that are assigned to a main rank in each of the five taxonomies. We found that all taxonomies have 1–2% of nodes with an intermediate rank ('sub-', 'super-' and other), except for Greengenes. Nodes with no rank assignment are found only in OTT (3.3%) and NCBI (13.3%). The latter taxonomic classification has the lowest percentage (84.4%) of nodes that fall into the category of main ranks.

Figure 2b shows the composition of the five taxonomies at all ranks down to the level of genus. The NCBI taxonomy has 2.7 times fewer genera and 1.9 times fewer species (not shown) than the OTT. In the following we describe each of the five taxonomies in more detail (summarized in Table 1).

SILVA

The SILVA database [4] contains taxonomic information for the domains of Bacteria, Archaea and Eukarya. It is based primarily on phylogenies for small subunit rRNAs (16S for prokaryotes and 18S for Eukarya). Taxonomic rank information for Archaea and Bacteria is obtained from *Bergey's Taxonomic Outlines* [10–13] and from the *List of Prokaryotic Names with Standing in Nomenclature* (LPSN) [14], whereas eukaryotic taxonomy is based on the consensus views of the *International Society of Protistologists* [15, 16]. Taxonomic rank assignments in the SILVA database are manually curated [4]. For the comparisons we used the taxonomy associated with SILVA small

subunit ribosomal RNAs (16S/18S) v128 as released on 29/09/2016.

Ribosomal database project (RDP)

The RDP database [3] is based on 16S rRNA sequences from Bacteria, Archaea and Fungi (Eukarya). It contains 16S rRNA sequences available from the *International Nucleotide Sequence Database Collaboration* (INSDC) [17] databases. Names of the organisms associated with the sequences are obtained as the most recently published synonym from *Bacterial Nomenclature Up-to-Date* [18]. Information on taxonomic classification for Bacteria and Archaea is based on the taxonomic roadmaps by *Bergey's Trust* [19] and LPSN [14]. Taxonomic information for fungi is obtained from a hand-made classification dedicated to fungal taxonomy [3]. For the comparisons we used a taxonomy associated with RDP database of 16S rRNA (Bacteria and Archaea) and 28S rRNA (Fungi) sequences as released on 30/09/2016 (release 11.5).

Greengenes (GG)

The Greengenes taxonomy [6] is dedicated to Bacteria and Archaea. Classification is based on automatic *de novo* tree construction and rank mapping from other taxonomy sources (mainly NCBI). Phylogenetic tree is constructed from 16S rRNA sequences that have been obtained from public databases and passed a quality filtering. Sequences are aligned by their characters and secondary structure and then subjected to tree construction with FastTree [20]. Inner nodes are automatically assigned taxonomic ranks from NCBI supplemented with previous version of Greengenes taxonomy and CyanoDB [21]. For the comparisons we used a taxonomy associated with the Greengenes database as released on May 2013. Although Greengenes is still included in some metagenomic analyses packages, for example QIIME [22], it has not been updated for the last three years.

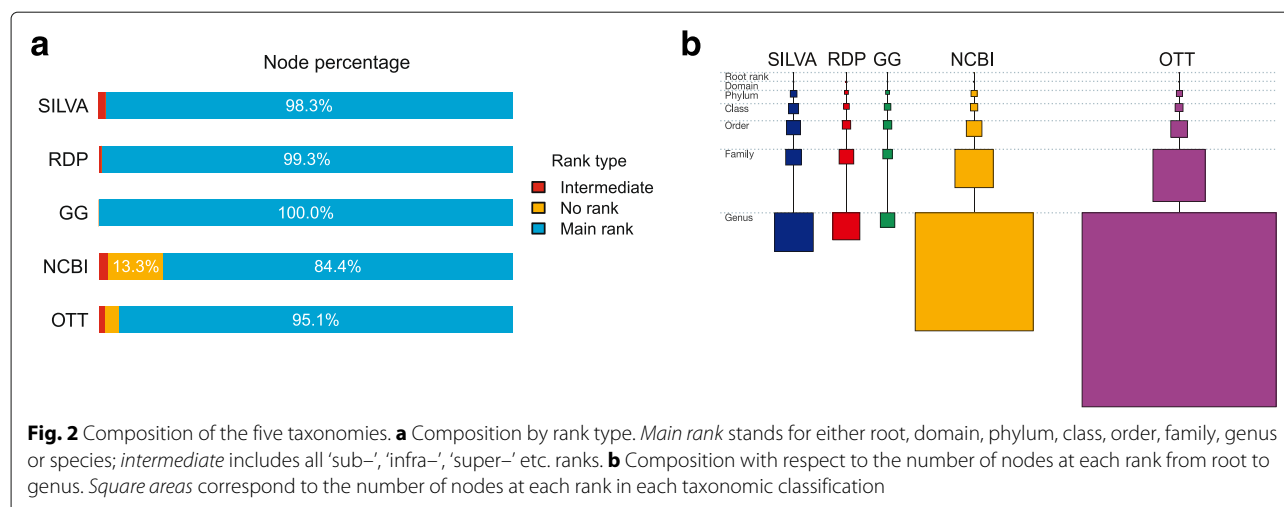
NCBI

The NCBI taxonomy [7] contains the names of all organisms associated with submissions to the NCBI sequence databases. It is manually curated based on current systematic literature, and uses over 150 sources, for example, the *Catalog of Life* [23], the *Encyclopedia of Life* [24], *Name-Bank* [25] and *WikiSpecies* [26] as well as some specific

Table 1 Overview of five taxonomic classifications

Taxonomy	Type	No. of nodes	Lowest rank	Latest release
SILVA	Manual	12,117	Genus	Sep 2016
RDP	Semi	6,128	Genus	Sep 2016
Greengenes	Automatic	3,093	Species	May 2013
NCBI	Manual	1,522,150	Species	Today ^a
OTT	Automatic	2,627,066	Species	Sep 2016

^aFor the analyses we have used NCBI taxonomy as published on 5th Oct 2016



databases dedicated to particular groups of organisms. It contains some duplicate names that represent different organisms. Each node has a *scientific name* and may have some *synonyms* assigned to it [7]. NCBI taxonomic classification files are updated on a daily basis; in this paper we use the version as of 05/10/2016.

Open tree of life taxonomy (OTT)

The Open Tree of life Taxonomy [9] aims at providing a comprehensive tree spanning as many taxa as possible. OTT is an automated synthesis of published phylogenetic trees and reference taxonomies. Phylogenetic trees have been ranked, aligned and merged together; taxonomies have been used to fill in the sparse regions and gaps left by phylogenies. Phylogenetic trees for the synthesis are obtained from TreeBASE [27], Dryad [28] and in some cases directly from contributing authors. Taxonomies are sourced from *IndexFungorum* [29], SILVA, NCBI, *Global Biodiversity Information Facility* [30], *Interim Register of Marine and Nonmarine Genera* [31] and some clade specific resources [9]. For the comparisons we used OTT taxonomy v2.10 draft 11 as generated on 10/09/2016.

Methods

Shared taxonomic units

First we determined how similar taxonomies are to each other by counting how many taxa they have in common at each rank. Similar comparisons have been carried out by Yilmaz et al. [4], however they confined their comparison to 16S databases, that is, SILVA, RDP and Greengenes; and only to phylum and genus levels.

We compared the number of shared taxonomic units (by name) between the four taxonomies: SILVA, RDP, Greengenes and NCBI, at each rank from phylum to genus. We then compared the union of the four taxonomies (ALL in Fig. 3) against the OTT in the same way (See

Fig. 3). To avoid differences because of alternative names or misspellings, we used the NCBI synonym dictionary to correct all names to scientific names.

Mapping procedure

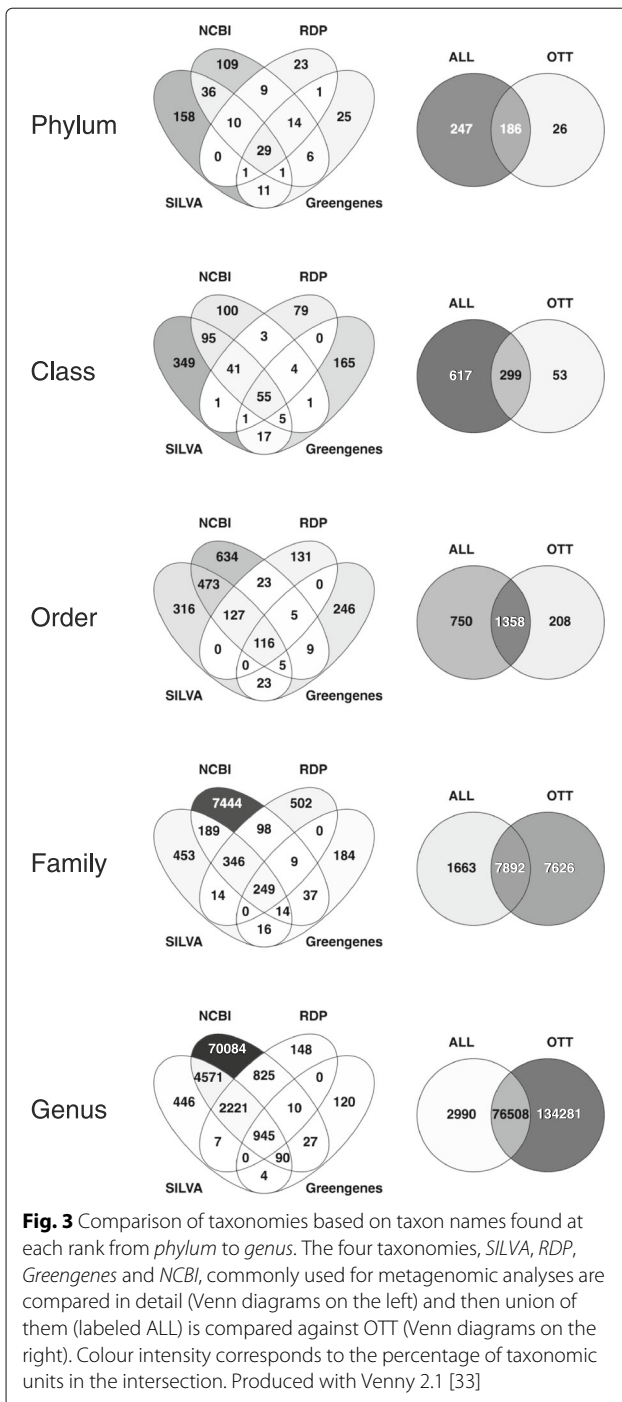
We define a procedure for mapping the nodes of one taxonomic classification onto nodes of another that is based on their hierarchical rank structure. As mentioned above, some of the taxonomies do not contain intermediate ranks, so we limit our comparisons to the seven main ranks. To get a detailed picture of how similar different pairs of taxonomies are, we perform three different mappings: *strict*, *loose* and *path comparison*, as we describe in the following.

Let A be source taxonomy that we are mapping into a target taxonomy B . Let $rank(a)$ be a function that defines the rank of a node $a \in A$ and $name(a)$ be the name of a . We say that we can map $a \in A$ perfectly into B when there is a node $b \in B$ such that $rank(a) = rank(b)$ and $name(a) = name(b)$. We denote a mapping of a as $\mu(a)$.

Let A and B be the two taxonomies to be compared. In the following we will assume that both taxonomies contained only nodes that are assigned to one of the seven main levels. To achieve this, we preprocess each taxonomy by contracting all edges that lead to a node that is not assigned to one of the seven main ranks, thus removing all such nodes.

Strict mapping

A *strict mapping* is calculated in a pre-order traversal as follows. If some node a has no perfect match in B , then we map a and all of its descendants to the same node as the parent of a . Note that we can always map the *root* node perfectly. See Fig. 4a for an example of a strict mapping on a set of nodes on a single path from root to species *Persicus* in Greengenes onto SILVA.



Loose mapping

A *loose mapping* is calculated in a pre-order traversal as follows. If some node $a \in A$ maps perfectly to a node $b \in B$ then we set $\mu(a) := b$. Let $a' \in A$ be a node that has no perfect mapping in B and a'' be an ancestral node of a' , then we map a' to the same node as a'' , i.e., $\mu(a') := \mu(a'')$. The main difference between the two kinds of mappings is that for the loose mapping, if

a is mapped perfectly to b , then we do not require that all ancestral nodes of a are also mapped perfectly (see Fig. 4b).

Path comparison

Path comparison is a special case of the loose mapping procedure. Here we take into account only those nodes in A that, themselves, or whose descendants, can be mapped perfectly onto B . In other words, we compare the paths from the root to the nodes with the same name and the same rank (see Fig. 4b).

Summary dissimilarity

By applying either the strict or the loose mapping procedure, each node $a \in A$ is mapped to some node $b \in B$. If the mapping is not perfect, then we will express this using a score based on the rank differences between such nodes a and b . To this end, we define the level of a rank as the distance from the root of taxonomy, that is, $level(root) = 0, level(domain) = 1, \dots, level(genus) = 6$. We ignore nodes at the rank of species because the *RDP* and *SILVA* taxonomies do not contain them, whereas *NCBI* and *OTT* both have more species nodes than the others have nodes in total.

The mapping distance for node $a \in A$ mapped to $b \in B$ is defined as $|level(a) - level(b)|$, a measure of how different A and B are with respect to placement of the node a .

The overall dissimilarity of two taxonomies A and B is calculated as the sum of all distances normalized by the sum of maximum distances:

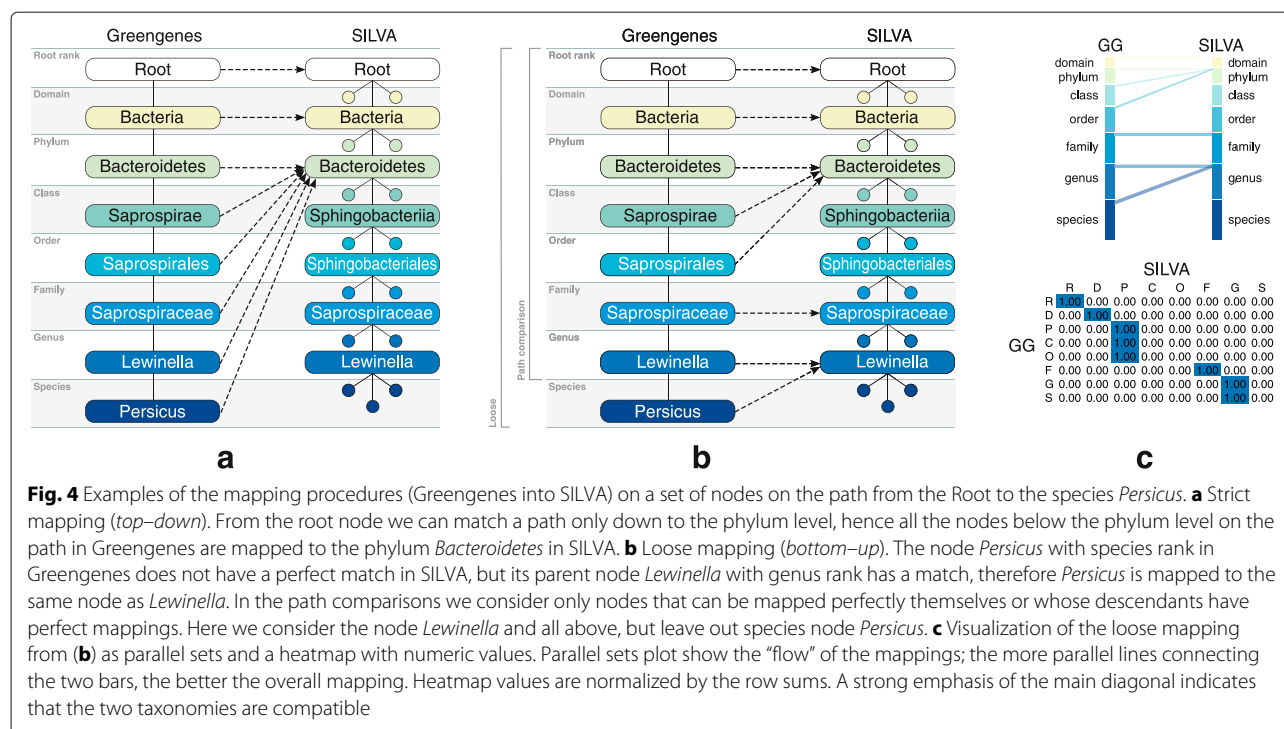
$$Q(A, B) = \frac{\sum_a (level(a) - level(\mu(a)))}{\sum_a level(a)}, \quad (1)$$

with the sum running over all nodes a with $rank(a) \neq species$ and $\mu(a)$ denoting the node in B to which a is mapped to. Note that $level(a) \geq level(\mu(a))$ for all $a \in A$. The value of $Q(A, B)$ lies between 0, indicating that all nodes in A are mapped perfectly to B , and 1, indicating that all nodes in A are mapped to the *root* of B , respectively. Note that the mapping dissimilarity is not symmetric, that is, in general we have $Q(A, B) \neq Q(B, A)$.

To allow a more detailed insight into the mappings, we provide summaries of the mapping results both as heat maps and as parallel sets (Fig. 4c). We provide such visualizations for all pairs of taxonomies and all types of mapping procedures in the Additional file 1. In the “Results” section we provide an overview graph of summary dissimilarities for all pairs.

Implementation

The mapping procedures described in this paper are implemented in a Java program called *CrossClassify*, which reads and writes taxonomic classifications in



BIOM1 [32] format. The program is Open Source (AGPL license) and is available from <http://ab.inf.uni-tuebingen.de/software/crossclassify/>.

Results

Comparison by shared taxonomic units

The simple comparison reported in Fig. 3 clearly shows that there are a lot of differences between the four taxonomic classifications, but there is also a lot of overlap, too. Each taxonomy at each rank has many taxa not shared with any other taxonomy – 73% of phyla, 70% of class, 63% of order, 90% of all families and 89% of all genera are unique to either SILVA, RDP, Greengenes or NCBI (OTT excluded). The NCBI taxonomy shares many more taxa with SILVA (60% in phylum, class and order ranks, and 10% in family and genus ranks) than it does with RDP (23% and 5%) or Greengenes (13% and 2%).

Interestingly, there are not many taxonomic units in the intersections that exclude NCBI (6% of phyla, 3% class, 1% order and < 1% of families and genera), indicating that the other three taxonomies are mostly contained in the NCBI taxonomy. Comparing the unions of the four taxonomies against OTT, an immediate conclusion is that there is more variety in the union (ALL) at the phylum (54%) and class (64%) ranks, and more in OTT at the genus (63%) and species ranks (not shown), whereas at order and family ranks highest portion of taxonomic units is shared among the union (ALL) and OTT (59% and 46% accordingly).

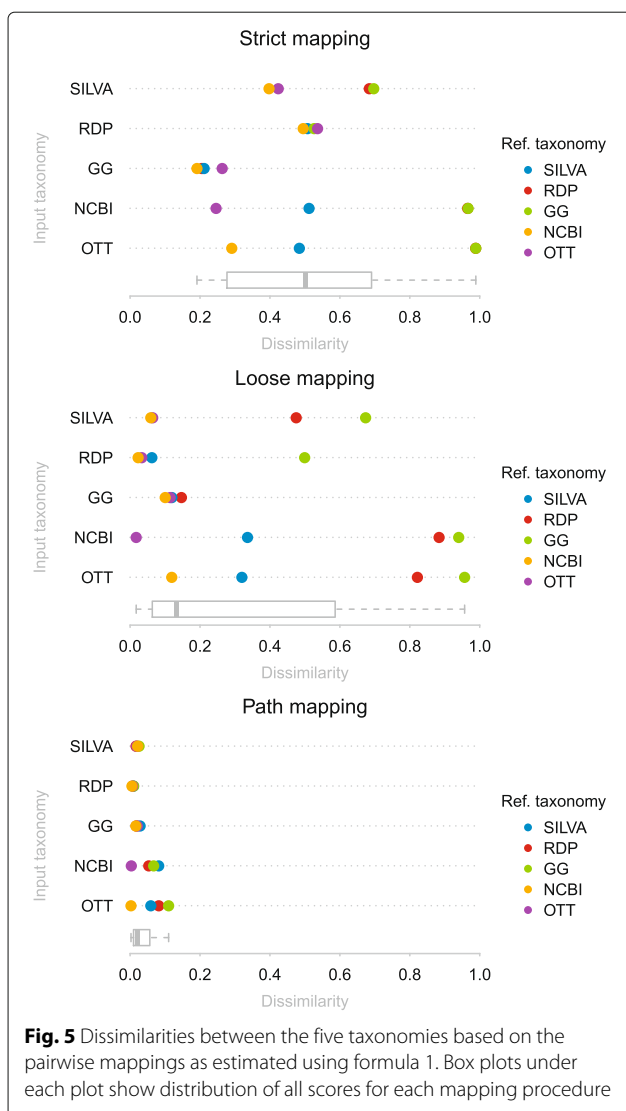
Comparison by mapping

We compare the five taxonomic classifications by mapping them onto each other using the three methods defined above. In Fig. 5 we show mapping scores for all pairs of taxonomies.

The strict mapping procedure gives very poor scores for most of the pairs with median dissimilarity of 0.5. Dissimilarities lower than 0.28 (25th percentile) are observed only for the mappings of Greengenes onto other taxonomies and for the mapping of NCBI onto OTT.

The loose mapping algorithm allows one to map nodes to closer ranks when possible and this is reflected in the dissimilarity distribution (median dissimilarity 0.13) as shown in Fig. 5. Loose mapping identifies RDP and Greengenes as the most difficult to map to with average dissimilarity of 0.58 for mappings on RDP and 0.77 for Greengenes. Loose mappings onto SILVA taxonomy have an average dissimilarity of 0.21 which is much better than for RDP and Greengenes, but not as good as for the two largest taxonomies – NCBI and OTT which have average mapping dissimilarities of 0.08 and 0.06 accordingly. However when mapping NCBI and OTT onto other taxonomies we get much worse average dissimilarities of approximately 0.68 for both. In fact, the NCBI taxonomy maps much better onto OTT (dissimilarity of 0.02) than vice versa (0.12).

The mapping of the common paths produces much lower dissimilarities (median 0.02) than the loose mapping procedure, albeit not perfect. That is, there is



some disagreement between the taxonomies on the paths to the same taxonomic units.

Other applications

The scoring functions and visualization techniques introduced in this paper to compare taxonomies can be used for other purposes, too. For example, they can be used to compare the behavior of two different taxonomic binning methods on the same set of sequencing reads and (same) taxonomy. In this context, the mapping function is defined by the two taxonomic assignments of each read. In a parallel sets plot, we scale the bars so that each bar is proportional to the number of reads that are mapped to the corresponding rank. For example, in Fig. 6 we display a comparison of the naive *Lowest Common Ancestor* (LCA) algorithm and the so-called *weighted LCA* (wLCA) algorithm [1], clearly showing that one method produces more specific taxonomic assignments than the other.

Discussion and conclusions

Comparing taxonomies by shared taxonomic units as shown in Fig. 3, we find that the number in the intersections is strongly limited by the size of the smallest taxonomy, Greengenes (see Table 1). SILVA, being the largest of the three 16S based taxonomies, shares the most taxonomic units with NCBI. We find that results of these simple comparisons are dominated by the number of nodes in each taxonomy and they tell us very little about structural (topological) compatibility of the taxonomic classifications. We address this issue by mapping taxonomies onto each other.

Our strict mapping procedure indicates how compatible the cores of taxonomic classifications are. Loose mapping on the other hand, has a less conservative nature and is closer to the comparison of shared taxonomic units. It indicates overall compatibility between taxonomies disregarding discrepancies at higher ranks that appear to be quite common; the median mapping dissimilarity for strict mapping is almost four (3.76) times as large as for the loose mapping, as shown in the box plots in Fig. 5.

The large difference between strict and loose mapping dissimilarities for NCBI to OTT (Fig. 5 and Additional file 1: Table S1) indicates that there are a few nodes with high rank that are incompatible between NCBI and OTT, but overall the two taxonomies are very similar (see Additional file 1 for more details). A much worse dissimilarity for loose mapping of OTT to NCBI (0.12) is most likely due to the fact that OTT has almost twice as many nodes as does NCBI.

Small differences (<0.05) between strict and loose mapping dissimilarities are mostly observed for the pairs of taxonomies where both dissimilarities are relatively high (>0.5). This is the case for all mappings on Greengenes. Again, this is not surprising since Greengenes is the smallest taxonomic classification and such differences indicate that it is much less diverse than other taxonomies. Dissimilarity for the loose mapping of OTT to NCBI is quite small (0.12) indicating that even though OTT is twice as large as NCBI, it is not much more diverse. This observation is also supported by differences in the number of nodes at each rank (Fig. 2, Additional file 1: Table S2) — the numbers for NCBI and OTT are very similar up to the family rank and differ significantly from genus rank only.

Path comparison dissimilarities indicate the scale of differences among paths to the taxonomic units shared pairwise by these taxonomies. Ideally all paths would be the same and their dissimilarities equal to 0. However, in this case there should also be no difference between the results of strict and loose mappings. Path comparisons show exactly how much difference there is in “shared” structure. Results (as shown in Fig. 5) show the same trend as discovered above – SILVA, RDP and Greengenes map

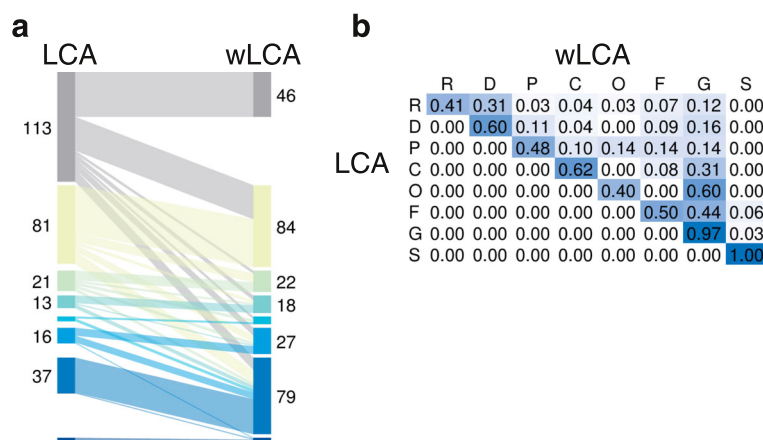


Fig. 6 Difference between taxonomic assignment with LCA and weighted LCA. Both plots indicate more specific assignments by weighted LCA as compared to LCA. Bars in the parallel sets plot in **a** correspond to the ranks from top as follows: root, domain, phylum, class, order, family, genus and species. Columns and rows in the heatmap in **b** correspond to the same ranks: R (root), D (domain), P (phylum), C (class), O (order), F (family), G (genus) and S (species)

well on NCBI and OTT but not vice versa. NCBI and OTT both are very similar with respect to path comparisons (dissimilarities are < 0.01).

Because OTT is the largest of the five taxonomies and because all other taxonomies map well on it, one might consider using OTT as the reference taxonomy of choice. However, at present OTT has no sequence database associated with it, which limits its usefulness in the context of metagenomics. Therefore, we recommend using the NCBI taxonomy as a common framework when comparing analyses performed on different taxonomic classifications. While the SILVA taxonomy is widely used for 16S studies, one should consider using the NCBI taxonomy in studies that use both targeted 16S sequencing and shotgun sequencing.

Additional file

Additional file 1: Supplementary material. A PDF file containing supporting data for the figures and detailed visualizations of pairwise mappings. (PDF 197 kb)

Abbreviations

GG: Greengenes; LCA: Lowest common ancestor; NCBI: National Center for Biotechnology Information; OTT: Open tree of life taxonomy; RDP: Ribosomal database project; wLCA: Weighted LCA

Acknowledgements

Not applicable.

Declarations

This article has been published as part of *BMC Genomics* Volume 18 Supplement 1, 2017. Selected articles from the 15th Asia Pacific Bioinformatics Conference (APBC 2017): genomics. The full contents of the supplement are available online <http://bmcbgenomics.biomedcentral.com/articles/supplements/volume-18-supplement-1>.

Funding

Publication costs were funded by Deutsche Forschungsgemeinschaft and Open Access Publishing Fund of University of Tübingen. MB was supported by a Teach@Tuebingen scholarship.

Availability of data and materials

The SILVA taxonomy analyzed during the current study is available in the SILVA file repository, [https://www.arb-silva.de/no_cache/download/archive/release_128/Exports/taxonomy/\(tax_slv_ssu_128.txt\)](https://www.arb-silva.de/no_cache/download/archive/release_128/Exports/taxonomy/(tax_slv_ssu_128.txt)). The RDP taxonomy analyzed during the current study is available in the RDP resources, <https://rdp.cme.msu.edu/misc/resources.jsp> (unaligned Fasta files). The Greengenes taxonomy analyzed during the current study is available in the Greengenes webpage, <http://greengenes.secondgenome.com/downloads>. The NCBI taxonomy analyzed during the current study is available in the NCBI ftp, [ftp://ftp.ncbi.nlm.nih.gov/pub/taxonomy/\(taxdmp.zip\)](ftp://ftp.ncbi.nlm.nih.gov/pub/taxonomy/(taxdmp.zip)). The OTT taxonomy analyzed during the current study is available in the Open Tree of Life webpage, <https://tree.opentreeoflife.org/about/taxonomy-version/ott2.10> (taxonomy.tsv).

Authors' contributions

DH proposed the project and suggested the mapping procedures. MB designed and implemented the code and carried out the comparisons. MB wrote the manuscript with some input from DH. Both authors read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Consent for publication

Not applicable.

Ethics approval and consent to participate

Not applicable.

Published: 14 March 2017

References

1. Huson DH, Beier S, Flade I, Górski A, El-Hadidi M, Mitra S, Ruscheweyh HJ, Tappu R. MEGAN Community Edition - interactive exploration and analysis of large-scale microbiome sequencing data. *PLoS Comput Biol*. 2016;12(6):1004957. doi:10.1371/journal.pcbi.1004957.
2. Pruesse E, Peplies J, Glöckner FO. Sina: Accurate high-throughput multiple sequence alignment of ribosomal RNA genes. *Bioinformatics*. 2012;28(14):1823–9.

3. Cole JR, Wang Q, Fish JA, Chai B, McGarrrell DM, Sun Y, Brown CT, Porras-Alfaro A, Kuske CR, Tiedje JM. Ribosomal database project: data and tools for high throughput rRNA analysis. *Nucleic Acids Res.* 2014;42(Database issue):633–42.
4. Yilmaz P, Parfrey LW, Yarza P, Gerken J, Priesse E, Quast C, Schweer T, Peplies J, Ludwig W, Glöckner FO. The SILVA and "All-species Living Tree Project (LTP)" taxonomic frameworks. *Nucleic Acids Res.* 2014;42(Database issue):643–8.
5. Wang Q, Garrity GM, Tiedje JM, Cole JR. Naïve Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Appl Environ Microbiol.* 2007;73(16):5261–7. doi:10.1128/AEM.00062-07.
6. McDonald D, Price MN, Goodrich J, Nawrocki EP, DeSantis TZ, Probst A, Andersen GL, Knight R, Hugenholtz P. An improved greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. *ISME J.* 2012;6(3):610–8.
7. Federhen S. The NCBI taxonomy database. *Nucleic Acids Res.* 2012;40(D1):136–43.
8. Beiko RG. Microbial malaise: How can we classify the microbiome? *Trends Microbiol.* 2016;23(11):671–9.
9. Hinchliff CE, Smith SA, Allman JF, Burleigh JG, Chaudhary R, Coghill LM, Crandall KA, Deng J, Drew BT, Gazis R, Gude K, Hibbett DS, Katz LA, Laughinghouse HD, McTavish EJ, Midford PE, Owen CL, Ree RH, Rees JA, Soltis DE, Williams T, Cranston KA. Synthesis of phylogeny and taxonomy into a comprehensive tree of life. *Proc Natl Acad Sci USA.* 2015;112(41):12764–9.
10. Boone DR, Castenholz RW, Garrity GM, Stanley JT, (eds). *Bergey's Manual of Systematic Bacteriology, Volume One: The Archaea and the Deeply Branching and Phototrophic Bacteria.* New York: Springer-Verlag; 2001.
11. Brenner DJ, Krieg NR, Garrity GM, Staley JT, (eds). *Bergey's Manual of Systematic Bacteriology, Volume 2: The Proteobacteria, 2nd edn.* US: Springer-Verlag; 2005.
12. Vos PD, Garrity GM, Jones D, Krieg NR, Ludwig W, Rainey FA, Schleifer KH, Whitman WB, (eds). *Bergey's Manual of Systematic Bacteriology, Volume 3: The Firmicutes, 2nd edn.* New York: Springer-Verlag; 2009.
13. Krieg NR, Staley JT, Brown DR, Hedlund BP, Paster BJ, Ward NL, Ludwig W, Whitman WB, (eds). *Bergey's Manual of Systematic Bacteriology, Volume 4: The Bacteroidetes, Spirochaetes, Tenericutes (Mollicutes), Acidobacteria, Fibrobacteres, Fusobacteria, Dictyoglomi, Gemmatimonadetes, Lentisphaerae, Verrucomicrobia, Chlamydiae, and Planctomycetes, 2nd edn.* New York: Springer-Verlag; 2010.
14. Parte AC. LPSN—list of prokaryotic names with standing in nomenclature. *Nucleic Acids Res.* 2014;42(Database issue):613–6.
15. Adl SM, Simpson AG, Lane CE, Lukeš J, Bass D, Bowser SS, Brown M, Burki F, Dunthorn M, Hampl V, Heiss A, Hoppenrath M, Lara E, LeGall L, Lynn DH, McManus H, Mitchell EAD, Mozley-Stanridge SE, Parfrey LW, Pawłowski J, Rueckert S, Shadwick L, Schoch C, Smirnov A, Spiegel FW. The revised classification of eukaryotes. *J Eukaryot Microbiol.* 2012;59(5):429–93. doi:10.1111/j.1550-7408.2012.00644.x.
16. Adl SM, Simpson AG, Farmer MA, Andersen RA, Anderson OR, Barta JR, Bowser SS, Brugerolle G, Fensome RA, Fredericq S, et al. The new higher level classification of eukaryotes with emphasis on the taxonomy of protists. *J Eukaryot Microbiol.* 2005;52(5):399–451.
17. Cochrane G, Karsch-Mizrachi I, Takagi T, International Nucleotide Sequence Database Collaboration. The international nucleotide sequence database collaboration. *Nucleic Acids Res.* 2016;44(Database issue):48–50. doi:10.1093/nar/gkv1323.
18. Leibniz Institute DSMZ-German Collection of Microorganisms and Cell Cultures Germany. Prokaryotic Nomenclature Up-to-date. <http://www.dsmz.de/bacterial-diversity/prokaryotic-nomenclature-up-to-date>. Accessed 11 Oct 2016.
19. Bergey's Manual Trust. Bergey's Taxonomic Outlines. <http://www.bergeys.org/outlines.html>. Accessed 11 Oct 2016.
20. Price MN, Dehal PS, Arkin AP. Fasttree: Computing large minimum evolution trees with profiles instead of a distance matrix. *Mol Biol Evol.* 2009;26(7):1641–50. doi:10.1093/molbev/msp077.
21. Komárek J, Hauer T. CyanoDB.cz – On-line database of cyanobacterial genera. – Word-wide electronic publication. <http://www.cyanodb.cz>, Univ. of South Bohemia & Inst. of Botany AS CR. Accessed 11 Oct 2016.
22. Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK, Fierer N, Pena AG, Goodrich JK, Gordon JL, et al. Qiime allows analysis of high-throughput community sequencing data. *Nat Methods.* 2010;7(5):335–6.
23. Roskov Y, Abucay L, Orrell T, Nicolson D, Flann C, Bailly N, Kirk P, Bourgoin T, DeWalt RE, Decock W, De Wever A. Species 2000 & its catalogue of life. Species 2000: Naturalis, Leiden, the Netherlands. 2016. ISSN 2405-8858.
24. Parr CS, Wilson N, Leary P, Schulz K, Lans K, Walley L, Hammock J, Goddard A, Rice J, Studer M, et al. The encyclopedia of life v2: providing global access to knowledge about life on earth. *Biodivers Data J.* 2014;2:1079.
25. Marine Biological Laboratory. uBio NameBank. <http://www.ebi.ac.uk/miriam/main/collections/MIR:00000338>, Woods Hole, Massachusetts, USA. Accessed 11 Oct 2016.
26. Wikimedia Foundation. Wikispecies. <http://species.wikimedia.org>. Accessed 11 Oct 2016.
27. Sanderson MJ, Donoghue MJ, Piel W, Eriksson T. TreeBASE: a prototype database of phylogenetic analyses and an interactive tool for browsing the phylogeny of life. *Am J Bot.* 1994;81(6):183.
28. Dryad. Dryad Digital Repository. <http://datadryad.org>. Accessed 11 Oct 2016.
29. Index Fungorum Partnership. Index Fungorum. www.indexfungorum.org/. Accessed 11 Oct 2016.
30. Global Biodiversity Information facility. GBIF Backbone Taxonomy. <http://www.gbif.org>. Accessed 11 Oct 2016.
31. CSIRO. Interim Register of Marine and Nonmarine Genera (IRMNG). <http://www.cmar.csiro.au/datacentre/irmng/>. Accessed 11 Oct 2016.
32. McDonald D, Clemente J, Kuczynski J, Rideout J, Stombaugh J, Wendel D, Wilke A, Huse S, Hufnagle J, Meyer F, Knight R, Caporaso J. The Biological Observation Matrix (BIOM) format or: how I learned to stop worrying and love the ome-ome. *GigaScience.* 2012;1(1):7. doi:10.1186/2047-217X-1-7.
33. Oliveros JC. Venny. An interactive tool for comparing lists with Venn diagrams. <http://bioinfogp.cnb.csic.es/tools/venny/index.html>. Accessed 11 Oct 2016.

Submit your next manuscript to BioMed Central and we will help you at every step:

- We accept pre-submission inquiries
- Our selector tool helps you to find the most relevant journal
- We provide round the clock customer support
- Convenient online submission
- Thorough peer review
- Inclusion in PubMed and all major indexing services
- Maximum visibility for your research

Submit your manuscript at
www.biomedcentral.com/submit

